DOCUMENT RESUME

ED 394 323 FL 023 727

AUTHOR Sigott, Gunther
TITLE Quantifying Language Ability.
INSTITUTION Council for Cultural Cooperation, Strasbourg (France).
PUB DATE 31 Jan 96
NOTE 23p.; Paper presented at the Educational Research Workshop on the Effectiveness of Modern Language Learning and Teaching (Graz, Austria, March 5-8, 1996)
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communicative Competence (Languages); Foreign Countries; Language Proficiency; *Language Tests; Models; Second Language Instruction; *Second Language Learning; *Testing

ABSTRACT

This paper highlights central topics in language testing theory and practice that are relevant to the examination of modern language teaching and learning. Part 1 gives a time-lapse picture of the development of models of language competence in language testing, discusses the distinction between descriptive models and working models, and reviews the problem of distinguishing between underlying knowledge and performance. Part 2 deals with the operationalization of models in the form of tests, considering first the relationship between test method and the concept of communicative language testing and then addressing reduced redundancy testing and rating scales as measures of communicative competence. This second part also addresses the importance of reliability and validity, makes a case for standards of practice in language testing, and reviews language test equivalency across languages. Part 3 suggests four concrete implications for policy: all tests should be reexamined for validity in light of new language testing theory; testing should be matched up with communicative language teaching; language test batteries with crosslinguistic applicability should be constructed for all languages of the European Union; and research on the impact of tests on teaching should be carried out. (Contains 66 references.) (Author/NAV)

Council of Europe
Conseil de l'Europe

Strasbourg, 31 January 1996

DECS/Rech (96) 8
Or. Engl.

COUNCIL FOR CULTURAL CO-OPERATION

Educational Research Workshop on the effectiveness of modern language learning and teaching, Graz (Austria), 5-8 March 1996

QUANTIFYING LANGUAGE ABILITY

by

Dr. Günther SIGOTT
Assistant Professor, Universität Klagenfurt,
Department of English and American Studies,
Universitätsstrasse 65-67, A-9020 KLAGENFURT

QUANTIFYING LANGUAGE ABILITY

Abstract

The paper attempts to highlight central topics in language testing theory and practice which are relevant in the context of examining the effectiveness of modern language teaching and learning. It is structured into three parts. Part 1 gives a time-lapse picture of the development of models of language competence in language testing, discusses the distinction between descriptive models and working models, and finishes by discussing the problem of distinguishing between underlying knowledge and performance. Part 2 deals with the operationalisation of models in the form of tests. It first considers the relationship between test method and the concept of communicative language testing by asking which test method features make a test a communicative test and suggesting an answer via the notion of test authenticity. This is followed by a section on reduced redundancy testing, which lists different types of reduced redundancy tests and focusses on the C-Test as a measure of general language proficiency. The next section discusses rating scales as measures of communicative competence. Shortcomings in rating scale validation are pointed out and suggestions for research in this area are made. The following section deals with language test equivalence across languages. It points out the usefulness of rating scales for this purpose but also attempts to make clear the limitations of this approach. The main problems inherent in test translation are discussed and some preliminary research results from studies involving translationally equivalent reduced redundancy tests are presented. The need for proficiency measures with crosslinguistic applicability as a prerequisite for studying foreign language learnability is pointed out. The last section highlights the importance of reliability and validity and makes a strong case for standards of practice in language testing. Part 3 suggests four concrete implications for policy: all tests currently in use or in the process of development should be examined for validity now that language testing theory has provided a more sophisticated theoretical basis for validation and the results of such studies should be publicised (1); testing should be matched up with communicative language teaching now that language testing theory is beginning to be able to accommodate the communicative movement also from a theoretical point of view (2); language test batteries with crosslinguistic applicability should be constructed for all languages of the European Union (3); and research into washback, that is, the impact of tests on teaching, should be carried out (4).

QUANTIFYING LANGUAGE ABILITY

Until recently, programme evaluation has been based on experimental designs using language test scores as dependent variables. Nowadays, however, there is an increasing awareness of the limitations of this approach, and so-called qualitative approaches are gaining importance. The evaluator of the 1990s has at his disposal a spectrum of methodologies ranging from quantitative to qualitative and eclectic approaches (Beretta 1992:18). Nevertheless, as with any new development, we have to be careful not to put all our hopes in the new and discard the old, but consider the new methodologies as supplements to the old, traditional, quantitative approach. This paper therefore focuses on the quantifiability and quantification of *linguistic* outcomes and on the quantification of language ability in general.

Language testing is a large and complex field which cannot be adequately covered by discussing a single research project in depth in a context like the present one. Language testing is characterised by a multiplicity of approaches, which is a consequence of the complexity of language itself. However, this paper does not provide a full-scale state of the art account of the field. Readers interested in such articles may wish to consult Alderson (1991a), Bachman (1991) and Skehan (1988, 1989, 1991). Rather, I attempt to highlight central topics in language testing which I consider relevant in a discussion of the effectiveness of modern language learning and teaching.

1 Knowing a language

Measuring psychological attributes presupposes a theory of the construct to be quantified. Among all the psychological constructs which have been the object of more or less successful attempts at quantification, language ability is doubtless the most complex. Describing it in all its aspects and ramifications is the central issue in modern linguistics. However linguists' aims in describing language are many and varied, with quantification ranking rather low on the list of priorities. Hence it comes as no surprise that answers to obvious questions which arise when one attempts to quantify language ability, are not always easily found in the linguistics literature.

In everyday terms knowing a language means being able to understand utterances in that language and being able to express content in that language appropriately. The most basic concern of language testing theory is to specify this ability, thus providing a theoretical basis for the construction of language tests and for the interpretation of language test results. In fact, there can be no theoretically sound measure of language ability unless what is being quantified is, or at least can be, properly defined. Linguistics, and in particular theoretical

linguistics, one of the most important potential feeder disciplines for language testing, has, because of its emphasis on discrete elements of language and its failure to pay due attention to the relationship and interaction among the individual levels of description, not been very helpful in defining the construct to be measured. In language testing, however, we have to face the complexity of language if we are to make relevant statements about people's degree of mastery of language as a whole or large portions of it. The language tester has to infer underlying ability from actual behaviour and to generalise from a sample of manifestations to a whole domain. Hence it comes as no surprise that comprehensive models of language ability were developed within language testing theory itself rather than in linguistics. The following sections outline the development of such models in language testing.

1.1 Early models

Structural linguistics provided the basis for the first generation of competence models, which described language knowledge in terms of levels-by-skills matrices. Typically, language was divided up into the levels phonology/orthography, morphology, vocabulary and syntax, which manifest themselves in the four skills listening, speaking, reading and writing (e.g. Lado 1961; Harris 1969; Heaton 1975; for summary see Oller 1979:173ff.). Models of this type, occasionally broadened to encompass fluency and aspects of sociolinguistics, were in use until the early 1980s, when the increasing influence of pragmatics led to the communicative movement, which in turn resulted in greater attention being paid to aspects of competence beyond the sentence level. Accordingly, the models were now said to describe *communicative* competence.

1.2 Canale & Swain (1980)

Canale and Swain's model of communicative competence represents an important elaboration on first-generation models by postulating three major components of communicative competence: grammatical competence (as described in earlier models), sociolinguistic competence, and strategic competence. Grammatical competence encompasses pronunciation and spelling, morphology, vocabulary, syntax, and word and sentence-level semantics. Sociolinguistic competence is described as the ability to choose appropriate speech functions as well as appropriate exponents for their surface realisation on the one hand and the mastery of rules governing cohesion and coherence in the sense of Halliday & Hasan (1976) on the other. While in the original 1980 formulation of the model, the latter aspect is subsumed under the cover term "sociolinguistic competence", Canale (1983) proposed discourse competence as a component in its own right. Thus, while sociolinguistic competence is responsible for appropriacy with regard to the extralinguistic context, discourse competence relates to appropriacy with regard to the linguistic context. Finally, strategic competence comprises abilities which enable the language user to compensate for breakdowns in communication which may be due to performance constraints or due to insufficient competence in any of the other areas.

1.3 Cummins

Cummins (1979) sees communicative competence as comprising two major components, which he refers to as Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP). Of these two, BICS are seen as the more basic

because everyone who has acquired a language under natural conditions possesses them. CALP, in contrast, only develops as the result of schooling. Cummins (1983) elaborates on this description by postulating that language proficiency can be characterised on two dimensions. One of them is defined by the extremes "context embedded" vs "context reduced communication". This dimension refers to the extent to which the speaker is able to perform without contextual aid, which is here seen as the possibility of meaning being negotiated between speaker and addressee. Context reduced communication, in contrast, is characterised by its reliance on linguistic form rather than on extralinguistic context to convey meaning. The other dimension, cognitive involvement, refers to how cognitively demanding an activity or task is. That is, it attempts to describe the "amount of information that must be processed simultaneously or in close succession by the individual in order to carry out the activity." (Cummins 1983:121)

1.4 Descriptive Models and Working Models

None of the models mentioned so far explicitly refers to the relationship among the aspects of competence that they describe. Neither the interaction among the components in actual language use nor the degree to which individual components can develop independently of the others is made explicit. Whereas the call for a detailed model of language use, incorporating as it would have to, assumptions about storage and processing, constitutes a big challenge for model building in the future, some progress has been made in understanding the degree of separability of components. Cziko (1984) has introduced a useful distinction between descriptive models and working models. Descriptive models, he argues, attempt to identify the What of language ability by postulating and defining, to different degrees of detail, the elements and rules, that is, the components of language ability. Working models, in contrast, go beyond descriptive models by focusing on the relationships between or among aspects of competence. Working models, unlike descriptive models, attempt to identify aspects of language ability which develop, or may develop, independently of each other, thus identifying factors of language ability. These factors, which, as the term suggests, derive from statistical analyses of test scores, may or may not coincide with components postulated in descriptive models.

In the late 1970s, numerous studies seemed to indicate that few, if any, aspects of language competence developed independently of one another. This claim was based on analyses of data from language test batteries and coincided with the advent of reduced redundancy tests like dictation and cloze. It was not long before the empirical findings met with an interpretation which related to reduced redundancy testing. The general factor which repeatedly resulted from factor analyses of language test data was interpreted as reflecting the workings of the learners' pragmatic expectancy grammar, which John Oller (1979:25) describes as follows:

> In the normal use of language, no matter what level or mode of processing we think of, it is always possible to predict partially what will come next in any given sequence of elements. The elements may be sounds, syllables,words, phrases, sentences, paragraphs, or larger units of discourse. The mode of processing may be listening, speaking, reading, writing, or thinking, or some combination of these. In the meaningful use of language, some sort of pragmatic expectancy grammar must function in all cases.

This view of language competence, which has since become known as the Unitary Competence Hypothesis (UCH), while not incompatible with descriptive models of the time, soon met with scepticism from an empirical point of view. Reanalyses of data which were originally used to adduce evidence in favour of the UCH showed that different factor solutions resulted if different variants of factor analysis were applied (for details see Skehan 1988:212). Thus, it was argued that the dimensionality of language proficiency is best represented by a general factor plus a few separate albeit correlating specific factors:

> [...] there is *both* a general language proficiency factor *and* a series of "divisible" factors of competence. The factors do not align themselves exactly with separate skills of speaking, listening, reading, writing, and grammar that might be postulated, but they show some correspondence with such skills. (Carroll 1983:91).

In the meantime, however, a more differentiated view has been taken. One the one hand, awareness of the factors capable of influencing statistical relationships has grown and more attention is being paid to nonlinguistic factors such as learning style and learning/acquisition environment (Cziko 1984), while on the other hand there are indications that the statistical structure of language proficiency changes from more unitary to more multifactorial and back to more unitary again as learners move from lower levels to higher levels of mastery (Sang et al. 1986, Weir 1995 on reading). More research is required in this area. Nevertheless, the notion of a unitary trait, which should perhaps better be termed general proficiency, cannot be thrown overboard, as Bachman (1991:673) concludes:

> The unitary trait has been replaced, through both empirical research and theorising, by the view that language proficiency is multicomponential, consisting of a number of specific abilities *as well as a general ability or set of strategies or procedures* (my italics).

At any rate, the influence of Cziko's distinction between descriptive models and working models is evident in the most widely discussed, if not accepted, model of language ability nowadays.

1.5 Bachman: Communicative language ability

Bachman (1990) postulates three main components of communicative language ability: language competence, strategic competence and psychophysiological mechanisms. Language competence is divided into organisational competence and pragmatic competence, each of which are further subdivided into different aspects of competence. Unlike in previous - purely descriptive - models, the structure of language competence here is hierarchical, which constitutes explicit claims about the relationships that hold among the subcompetences. Thus, the model claims, for example, that the development of syntactic competence is more strongly related to that of phonological competence than it is to aspects of textual competence such as cohesion or rhetorical organisation (coherence). While organisational competence is subdivided into grammatical competence and textual competence, pragmatic competence comprises illocutionary competence and sociolinguistic competence. Grammatical competence consists of vocabulary, morphology, syntax and phonology/graphology; textual competence of cohesion and rhetorical organisation. Illocutionary competence, broadly

speaking, consists of knowledge of different types of speech acts or speech functions and their possible realisations (exponents), while sociolinguistic competence describes the knowledge necessary to choose from several possible realisations a surface realisation (exponent) for a speech act or speech function which is appropriate to the individual speech situation.

Bachman's model of language competence, while making explicit claims about the relationships among subcompetences, has not been empirically validated in all its aspects. While the distinction between organisational competence and pragmatic competence seems to be empirically justified, the model is currently undergoing modifications with respect to more detailed aspects of competence. The most important of these is the move of vocabulary from grammatical competence to pragmatic competence, where it now figures as lexical knowledge at the same level as functional knowledge (previously illocutionary competence) and sociolinguistic knowledge (Bachman & Palmer, forthcoming).

In contrast to Canale and Swain, who see strategic competence as purely compensatory, Bachman's notion of strategic competence is more positively formulated. It characterises "the mental capacity for implementing the components of language competence in contextualised communicative language use" (1990:84). In the modified version of the model (Bachman & Palmer, forthcoming), it is seen as a metacognitive process mediating between language knowledge (formerly language competence), knowledge schemata (i.e. knowledge and experience of the world) and affective schemata (i.e. emotional memories), and involving assessment strategies, planning strategies and goal-setting strategies. It is heartening to see how model building for language testing purposes is linking up with developments in discourse analysis, where a view of discourse as process is becoming more and more widespread (Cook 1989:57ff.).

1.6 Knowledge and performance

While from the beginning of scientific language testing to the present day the models have become more comprehensive by including aspects of language beyond the sentence level as well as aspects of pragmatics, this development is parallelled by greater attention being paid to the ability which is necessary for making use of this underlying knowledge. In the early, first generation, models this ability does not figure at all, and if implicit in them, then it was presumably seen as inextricably related to or inherent in the four skills of listening, speaking, reading and writing. It took over a decade for the notion of ability for use (Hymes 1967) to be tentatively accommodated in a model of communicative competence (Canale & Swain 1980). Interestingly, Canale and Swain explicitly exclude ability for use from their model of underlying knowledge by claiming that there is no theory of human action that can adequately explicate this notion (1980:7). And yet the notion of ability for use figures in their model in the guise of strategic competence, which, however, is accorded an exclusively compensatory function (McNamara 1995:168). In the early 1980s, strategic competence is seen as the ability to deploy defensive strategies when underlying knowledge is lacking or extraneous factors such as noise or fatigue on the part of the speaker/writer or the hearer/reader interfere with communication. Canale (1983) attempts to rectify the inconsistency inherent in the original Canale and Swain model by claiming that "this notion of skill - how well one can perform knowledge in actual situations - requires a distinction between underlying capacities (competence) and their manifestation in concrete situations" (Canale 1983:6),

thus arguing for ability for use to be included in the model. But by postulating discourse competence as a fourth aspect of underlying knowledge, Canale adds to the inconsistency. Discourse competence, he claims, is "the mastery of how to combine grammatical forms and meanings to achieve a unified spoken or written text in different genres. ... Unity of a text is achieved through cohesion in form and coherence in meaning" (1983:9). However, whether the ability to create coherence in discourse is a question of underlying knowledge exclusively, as Canale would like to have it, is unclear (McNamara 1995:169). Rather, it seems, it would make sense to follow Widdowson (1979:ch.10) and distinguish rules from procedures in the interpretation of discourse, the latter being part of strategic competence and the former pertaining to language competence.

Clearly, all of this begs the question of whether strategic competence really is part of language knowledge at all. Bachman, while seeing it as part of communicative language ability, separates it from language knowledge by describing it as "a general ability, which enables an individual to make the most effective use of available abilities in carrying out a given task" (1990:106). And yet, like Canale's discourse competence, Bachman's illocutionary competence overlaps with strategic competence if, as Bachman claims, it is involved when it comes to processing "a sentence type whose form is not generally associated with the given illocutionary act, and whose interpretation depends very heavily on the circumstances under which the act is performed" (Bachman 1990:91).

The scope of strategic competence, then, has been extended from a largely defensive role to a capacity which is thought to be a prerequisite for smooth and efficient communication in general. Apart from the means of repairing breakdown in communication, it includes the capacities underlying processes of assessment, planning and goal-setting (Bachman & Palmer, forthcoming) as well as forms of routinised communication (Little 1994:8). This latter aspect is clearly related to the concept of fluency, which plays an important part in the assessment of communicative competence by means of rating scales.

While some degree of consensus has emerged with regard to the knowledge aspect, model building for the performance aspect has hardly begun. It is an area of prime importance, though, because it will provide a firmer theoretical basis for communicative language testing. A model of strategic competence is required if we are to be able to generalise from scores on so-called communicative tests to underlying abilities. The model is necessary for adequate sampling of test content on the one hand, which is one of the prerequisites for generalisability of test results, and it is necessary as a theoretical basis for understanding the influence of non-linguistic factors on test performance, which in turn is a prerequisite for the construction of fairer tests.

A model of strategic competence is also required for another reason. The emergence of the concept as a prerequisite for language use has been paralleled by attempts to describe strategies employed by learners which result in language acquisition. In fact, it seems difficult, if not impossible, to draw a clear dividing line between strategies of language use and strategies of language learning (Little 1994). Strategic competence would seem to comprise two components, namely strategic competence for use and strategic competence for learning. Bialystok (1990, 1991) has proposed a model of language processing which takes account of the inseparability of strategies of use from strategies of learning. It posits the two subskills of analysis and of control, which interact to form the basis of the language pro-

cessing ability. Analysis covers processes which analyse linguistic input and gradually build up a representational system of language knowledge. Control, on the other hand, refers to the ability to select the appropriate focus of attention (i), to integrate information from a multiplicity of sources (ii), and to operate within the constraints imposed by restrictions of time (iii) (Little 1994:13). Whenever language is used, processing takes place. However, in no instance of communicative language use can processing take place without at the same time triggering processes of analysis to some extent. Thus, the distinction between processing for use and processing for learning, although conceptually possible, should not be taken to imply that processing of language can be categorised as either exclusively analysis or exclusively control. It follows that strategies of use are closely connected to, if not to a certain extent identical with, strategies of learning. Clearly, these notions await concretisation in a model of strategic competence which would also have to clarify the relationship of the construct to such constructs as general intelligence or language aptitude.

1.7 Summary

Although the area of model building is in continuous flux, it seems safe to make the following statements:

- The Bachman model (and I daresay its modified version in the future) serves as the most important point of orientation for both language testing research and test development.

- Language ability is generally seen as comprising a static and a dynamic aspect, both of which are considered types of underlying ability and are referred to by means of dichotomies such as declarative-procedural, knowledge-skill or knowledge-ability. Strategic competence is now commonly used to denote the procedural aspect. Because strategic competence has not yet been modelled to the extent other aspects of competence have, the dividing lines between strategic competence and language competence on the one hand, and constructs like general intelligence or even language aptitude on the other, are still unclear.

- The problem of defining language ability can be broken down into a qualitative aspect (the WHAT of language ability), a relational aspect (how independent of each other are the individual aspects in the development of language ability), and a quantitative aspect (how important are the individual aspects as parts of the whole). Considerable progress has been made with regard to the qualitative aspect, as the widespread acceptance of the Bachman model shows. Some progress has been made with regard to the relational aspect, as the hierarchical structure of the Bachman model indicates. But nothing is known about the quantitative aspect, with the result that a theoretical basis for weighting subtests in a test battery is still missing.

- Although the notion of general proficiency still awaits more detailed description, particularly in terms of its relationship to strategic competence, the concept seems to be a valid one.

2 From model to test

In an ideal world all test construction would be based on a model of the abilities to be measured. However, reality has been different and tests have been developed without recourse to models, either because such models were not available or because practical concerns prevailed over theoretical considerations (for an historical account of the development of the discipline see Spolsky 1995). Even today, tests are sometimes developed on the basis of practitioners' intuitions about language, which, however, are not always made explicit. Nevertheless, it is important to note that a model of the abilities to be measured is implied in every test, no matter whether it is made explicit or even whether the test writer is able to do so.

This is not the place to provide a detailed account of the individual stages of test construction, a process which is described and discussed in detail in Alderson, Clapham & Wall (1995). Instead, the following discussion focusses on selected aspects which I consider important and which, to a certain extent at least, relate to my own research interests.

2.1 Test method and communicative language testing

For a long time, the way a test elicits performance was, quite understandably, almost exclusively a concern of test or item writers. Its importance as a determinant of the test score was only acknowledged slowly, and few studies addressed the question up until the late 1970s, when several authors noted the effect the test method had on test scores. Only recently has the importance of the test method been made explicit by listing it along with competence effects, personal attributes and measurement error as the main determinants of the test score. Even a framework for characterising test method is now available (Bachman 1990, ch.5). This constitutes an important step toward a theoretical basis for communicative language testing.

The term 'communicative', whether applied to teaching or to testing, is surrounded by a haze of confusion. In particular, it has been asked whether communicative testing means testing communicative language or testing language communicatively (Davies 1995; Rea-Dickins 1991). Is it perhaps more of a How, a method, than an approach that focusses on a different What, a more comprehensive underlying ability? This question is the object of much debate among language testers at the moment and it is impossible to predict what the consensus will be if it ever emerges at all.

At any rate, the Bachman framework of test method facets is a useful basis for asking questions about the relationship between test method and communicative language testing. The Bachman method framework is not an attempt at characterising only those aspects of language tests which focus on abilities other than language knowledge, as the term method might be taken to indicate. It is an instrument which does justice to the fact that language tests are operationalisations of models of underlying ability, but as such also incorporate aspects of performance conditions. It is tempting to argue that all the facets of the method framework will have counterparts in underlying ability in the form of component abilities. However, theoreticians disagree over the extent to which such a claim is tenable. What are, for example, component abilities corresponding to facets of the test rubric such as salience or sequence of parts or to facets of the input such as identification of problem or degree of

speededness (Bachman 1990:119)? In the Bachman paradigm, the answer is strategic competence, which, however, has so far defied explicit description.

However, it is premature to conclude that communicative tests are simply tests of strategic competence. Communicative testing has been defined via the notion of authenticity (Bachman 1990:301). Bachman (1991) distinguishes situational authenticity and interactional authenticity, both of which he describes with reference to the methods framework. A test is situationally authentic, he argues, if its characteristics are perceived as corresponding to the features of a target language use situation (1991:690). For example, if a test for engineers contains relevant specialised vocabulary and topics in the input, this will make for situational authenticity. The second aspect of authenticity, interactional authenticity, depends on the extent and type of involvement of test takers' language ability in accomplishing a test task. It is, therefore, a matter of coverage of the elements of language knowledge and the degree to which aspects of strategic competence such as assessment, goal setting and planning are engaged by the test task. It is, it seems, basically a question of how integrative (Carroll 1961) a test task is.

Defining communicative tests via the notion of authenticity, which in turn can be discussed in terms of method facets, means viewing 'communicative' as a relative concept which is audience and situation specific. Given the variety of language use contexts, a test may be communicative for one type of candidate but not for another. Clearly, this begs the question of how generalisable results from communicative tests are. McNamara (1995; in press) addresses the issue. Coming from Language for Specific Purposes or performance testing, he argues for a model of the abilities underlying communicative performance. Generalisability from test performance in one situation to performance in other situations crucially depends on the availability of such a model. This means carrying Bachman's attempt to its logical conclusion although it remains to be seen how rigorously the abilities underlying communicative performance, particularly those making up strategic competence, can be described. Perhaps it is helpful to have critical voices like Davies (1995), who warn against too much enthusiasm. Nevertheless, McNamara's interest in generalisability and hence in an ability underlying several different performance tests is symptomatic and "may [...] be reflecting the general trend, which appears to be returning to a more unitary view of language ability" (Davies 1995).

2.2 Reduced Redundancy Testing

As mentioned earlier, the use of cloze as a measure of overall language proficiency was coincidental with the emergence of the Unitary Competence Hypothesis. Something similar is happening at the moment with the C-Test, although just how coincidental with the modified view of the structure of language proficiency the growing popularity of the C-Test is could be debated. As mentioned in 1.4 above, the concept of language proficiency as being best represented by a general trait plus a number of specific abilities still seems to be a valid one. What this general trait consists of is, as it always has been, a matter of debate. One position is to consider it as representing the set of abilities which are tapped into by reduced redundancy tests.

The theoretical rationale underlying reduced redundancy tests has been a fixture in language testing research and practice for at least twenty years. The basic assumption is simple and is

rooted in information theory: A sender transmits a message over a channel to a receiver. Noise in the channel may blur the message. If this happens, the receiver may still manage to decode the message by making use of the redundancy that is built into the message. The more familiar the receiver is with the language system and its workings, the more blurring he or she will be able to tolerate and still understand the message. This rudimentary theory, if such it can be called, has been operationalised in a number of ways: as Dictation (where the noise can be seen as simulated by the absence of visual clues like punctuation and capital letters), as the Noise Test (where white noise is added to a spoken text at varying levels of intensity and the testee has to perform tasks based on the spoken text), as Partial Dictation (where phrases selected by the test constructor are deleted and have to be supplied by the testees), as Cloze (with random or rational deletion or multiple choice), as Cloze-Elide (where irrelevant words are inserted), and finally, as the C-Test, which in its original format (Raatz & Klein-Braley 1985) deletes the second half of every second word (for a summary description of these various operationalisations of the theory see Klein-Braley 1994:18ff.).

Due to undeniable advantages for testing practice (ease of construction, ease and objectivity of scoring - if only the original word is counted as correct and any other linguistically acceptable alternatives are not), the cloze format enjoyed considerable popularity in the late 1970s. However, a number of flaws were detected in its psychometric properties (Alderson 1979, 1980, 1983; Klein-Braley 1981). As a result, Raatz and Klein-Braley (1985) proposed the C-Test, which incorporates a number of advantages over the cloze format: more items are possible with much shorter texts; few items have more than one possible solution; scoring is effortless for a person with nativelike command of the language; the chance of damaging a representative sample of all word classes in the text is higher than in cloze, and, finally, the sampling of content classes is better since a C-Test consists of at least four different passages (Klein-Braley 1994:42f.). Reliability coefficients for internal consistency are generally higher than .8 and a considerable amount of research has been showing that the C-Test is a valid measure of global language proficiency (for an overview see Grotjahn 1992, 1994, in press). C-Tests are thus well suited as placement and proficiency tests. They can also be used as indicators of global learning progress in general and ESP language courses (Sigott, in preparation). But they are not suitable as diagnostic or achievement tests.

The format also lends itself to modification in order to increase the difficulty of passages to make them suitable for populations in which otherwise ceiling effects would be encountered. These modifications involve increasing the rate of redundancy reduction by varying the proportion of letters deleted. For native speakers of German, the difficulty of C-Tests can be increased by ca. 20 percentage points by increasing redundancy reduction, that is the proportion of letters deleted, from 28% to 36%. A further reduction of redundancy to 43% increases difficulty by another 15 percentage points (Köberl & Sigott 1994). For native speakers of English, the analogous figures are ca. 20 and 10 percentage points respectively (Sigott & Köberl, in press). For German-speaking beginning university students of English, the same modifications bring about an increase in difficulty by ca. 14 and 10 percentage points respectively (Sigott, in preparation). In none of the studies conducted so far have the modifications led to sacrifices in the strengths of the original C-Test format. Thus, it seems fair to conclude that the format can easily be adapted to the testees' level of proficiency by simply changing the rate of redundancy reduction.

Despite these attractions, the C-Test format suffers from a serious drawback: the ability which it measures has never been described in any detail. Precisely this is the reason for a certain amount of scepticism that surrounds its use. However, recent and current research gives rise to the hope that the ability can be concretised. This research (Klein-Braley 1994; Sigott, in preparation) studies the relationship between item difficulty and item type and is beginning to make out a hierarchy of micro-abilities corresponding to different types of C-Test items. This hierarchy corresponds to a presumptive scale ranging from low-level processes to high-level processes.

2.3 Rating scales

The adoption of communicative approaches to language teaching has also brought about a need for assessment instruments that are compatible with the newly formulated course objectives. The emphasis in communicative language teaching is on the integration of microskills in language use, hence on strategic competence, and on the authenticity of tasks. Moreover, there is a tendency to stress the productive skills. Whether this is justified is a different issue and need not concern us here. At any rate, the emphasis on the productive skills, particularly on speaking, brings about a need for measuring instruments other than paper and pencil tests.

The basic difference in the approach to quantification that is taken by means of rating scales is made explicit by Pollitt (1991), who distinguishes between counting and judging. Counting involves the use of item-based tests, whereas judging involves the use of rating scales. Clearly, rating scales were in use before Pollitt's article, the prototype probably being the Foreign Service Institute Oral Interview, which was used as early as the 1950s. However, the basic difference in the approach to quantification was not made explicit before. Whereas in counting, a score is arrived at by adding up observations, the use of rating scales involves comparing a stretch of performance with an internalised idea of adequate performance and deciding how far away from that adequate performance the actually observed performance is.

Should performance be rated on one single dimension or on several dimensions? It is tempting to try and answer this question on the basis of what is known about the structure of language proficiency. As was noted in 1.4 above, a general factor plus a few separate factors is now widely accepted as an adequate representation. But since the number of factors and their precise nature remain unclear, it is impossible on this basis to decide on the number and type of scales that are to be used. Moreover, practical considerations may override theoretical concerns. Many contexts of assessment require as the end product a single score. In such cases, holistic, as opposed to analytic, scales may be used to begin with. North (1993) surveys some of the theoretical issues underlying the construction of rating scales and provides a survey of rating scales which are in use in various contexts and, where available, presents justifications for the use of either holistic or analytic scales in individual situations.

It is difficult to orientate oneself in the great variety of scales that are in use today. Alderson (1991b) usefully distinguishes between user-oriented, assessor-oriented, and constructor-oriented scales. User-oriented scales are couched in simple, non-expert terminology which is understandable to test takers and to people like employers, who have to understand what a given score on a test means in terms of the candidate's likely performance in real-life situations. In fact, an important function of user-oriented scales is to serve as a reporting

device on which scores from test batteries can be placed in order to help outsiders understand what a given score on a test actually means. Assessor-oriented scales involve expert terminology and in their descriptors focus on salient features of performance at each level. Consistent interpretation of the terms is ensured through rater training. Constructor-oriented scales are aimed at test constructors and help them to draw up test specifications and develop tests which are appropriate for individual levels of proficiency.

If standards of practice (see below) are observed, rating scales should be subjected to the same procedures of test quality control as item-based tests. That is, information concerning validity and reliability should be made available. However, as far as reliability is concerned, research on rating scales tends to overemphasise the rater as a source of unreliability. Little attention is paid to test-retest reliability and internal consistency (Pollitt 1991). In fact, very little is known about the influence the task which candidates have to perform exerts on the quality of the performance which is elicited. Consequently, a central question is how stable and generalisable to other situations a given rating is. This is an area which deserves a lot more attention in future research.

Rating scales also raise another research issue. In many scales the level descriptors are formulated in terms which suggest an order of skill acquisition, thus implying the existence of a hierarchy of skills or stages of acquisition through which the learners progress in their linguistic development. Whether learners indeed follow the pattern of development that is implicit in various rating scales is a question which can only be answered by means of empirical studies. The issue is now increasingly attracting attention (e.g. Pienemann et al. 1988; Brindley, in press) and gives rise to the hope that cooperation between language testing research and language acquisition research will intensify in the future.

Thus, the constructs involved in rating scales call for closer scrutiny. This becomes very obvious in Fulcher's (1993, ms.) work on the dimension of fluency, who demonstrates that the relationship between observable behaviour and the learners' progression toward the target is not isomorphic. For example, Fulcher has demonstrated that the incidence of hesitation phenomena is not linearly related to learners' progress. Hesitation is frequent among low-level learners, decreases significantly at intermediary level, and is frequent again at a very advanced level. Moreover, the causes of hesitation are different at different levels of proficiency. Observations of this type clearly show that ratings based on observable behaviour may be seriously flawed as indicators of developing competence.

Despite a certain lack of empirical research to justify the widespread use of rating scales for a variety of purposes, they are an important tool in the language tester's arsenal in the 1990s. They enjoy great popularity in vocational language contexts (e.g. Languages Lead Body 1993) and wherever there is a concern with test and task authenticity. One of their attractions is their potential usefulness in comparing levels of proficiency across languages.

2.4 Testing across languages

The growing unification of Europe and the corresponding increase in communication among its peoples has led to a need for instruments for comparing proficiency across languages. But it is only recently that the question has been seriously addressed. Thorny problems inherent in the translation of language tests have dictated the approach that is being adopted. It

basically consists in developing frameworks of level descriptors which, because they are formulated in functional terms, can reasonably be translated and, in principle at least, be applied to any language. Thus, the Languages Lead Body provides a five-level framework for the four skills which is intended as a basis for course development on the one hand, and on the other hand as a system of reference for awarding national language-related vocational qualifications in Great Britain. The specifications and level descriptions in the National Language Standards (Languages Lead Body 1993), are phrased in English and are considered to be applicable to any language. Another attempt at achieving comparability of assessments in different languages is the Council-of-Europe based work on a Common European Framework (North 1993, 1994), which on the one hand aims at equivalence of assessment but on the other also attempts to provide a basis for the formulation of objectives in the languages of the European Union. Achieving comparability of assessment and certification in different languages is also the main aim of the Association of Language Testers in Europe's Framework (ALTE 1995a). This framework is partly based on the Council of Europe's definitions of learning objectives (van Ek 1975) and like the National Language Standards, comprises 5 levels, within which Tests and Certificates in 10 European languages are placed. At this point, the ALTE framework is probably the most user-orientated in that it uses simple, non-expert terminology to describe its levels. This is done in terms of Can-do statements which serve as indicators of what somebody who holds a certificate at any of the five levels can do in real life. There are now over 50 scales of Can-do statements for particular activities such as shopping; requesting work-related services; or following a lecture, talk, presentation or demonstration.

These frameworks are useful and they constitute our currently available tool for comparing proficiency across languages. Nevertheless, a five-point scale is a very coarse measure if it is to be used for the whole spectrum of language proficiency ranging from absolute beginner to nativelike control. Moreover, the method by which examinations are placed onto the five-level scale is intuitive and lacks firm empirical justification. If a higher degree of precision is to be achieved in our tools for comparing proficiency across languages, the question of test equivalence across languages needs to be addressed at a more concrete level. If comparisons within the same level of the framework are to be made or if examinations are to be anchored onto the common scale on an empirical basis, item-based tests have to be used which can be scored objectively. This raises the question of language test translatability, an area which research has barely started to explore.

In Sigott (1992) I attempt an analysis of the problem and report some empirical results concerning cloze tests for English and French. Establishing test equivalence across languages involves establishing equivalence with regard to content and equivalence with regard to difficulty. Equivalence may be achieved through translation or through sampling. Discrete point tests focussing on phonology, morphology, vocabulary or syntactic structures are hardly ever translatable, particularly not when they are in a multiple-choice format. If, however, sampling is resorted to, the approach is complicated by the absence of a competence model which provides information on the weights of the individual subcompetences, which may have to be different in different languages. In contrast, text-based, integrative tests are generally translatable although it remains uncertain to what extent translation of such a test automatically ensures equivalence with regard to difficulty. For instance, a carefully translated and piloted cloze passage behaved differently in terms of difficulty in English and in French, depending on the point of onset of deletions, the scoring method and the candidates'

level of proficiency (Sigott 1992:163f). This result should not be interpreted as proving the uselessness of cloze as a crosslinguistic proficiency measure. Rather, it points to the necessity of relating scores on translationally equivalent passages to stable native-speaker norms for these passages before candidates' scores are compared across languages. This, however, means calling into question a tacitly assumed requirement for test validity, namely that native speakers must be able to make perfect scores on a language test. If meaningful native speaker norms for tests are to be obtained, this requirement must be waived. Reduced redundancy tests hold potential as crosslinguistic proficiency measures because of their ease of construction and translatability. In a pilot study involving translationally equivalent C-Test passages in English and German (Sigott & Köberl, in press), the C-Tests in both languages yielded satisfactory reliability coefficients for matched native-speaking groups. However, modifications of the deletion pattern had differential effects in the two languages. In the original format (C25), which deletes the second half of every second word, the English test was more difficult than the German equivalent. Similar results were obtained for other modifications of the deletion pattern; one that damages two thirds of every second word (C33) and one that leaves only the first letter of every second word (CFL). By contrast, if the first, rather than the second, half of every second word was damaged (X25), the German test was considerably more difficult, which is presumably due to differences in the morphological structure of the two languages. These results demonstrate the importance of relating comparisons of proficiency across languages to norms for speakers with native-like command of the languages involved. Otherwise serious misinterpretation of score differences will result.

The availability of accurate and reliable crosslinguistic proficiency measures will make it possible to address a question which researchers have so far carefully avoided. It is the question of the magnitude of the learning task that a particular foreign language presents for speakers of a given native language. Practitioners' informal observations as well as a pilot study into the question for English and French in the Austrian secondary school context (Sigott 1993) indicate that measurable differences in learnability exist even among the languages of the European Union. If multilingual competence and efficiency of foreign language learning programmes are aimed at, then the learnability question becomes highly relevant. We need to ask how much learning time needs to be allocated to different foreign languages for speakers of a given native language if equal levels of proficiency are to be attained. Learnability should not be forgotten in a framework which considers factors governing the efficiency of foreign language learning and teaching. However, the issue needs a lot more research in the area of crosslinguistic proficiency measures and once such measures are available, international cooperation will be essential.

2.5 Quality Control

Language tests are operationalisations of models of competence. How well a test operationalises a model of competence is one of the most central questions of language testing theory and practice, namely that of validity. A test can only be valid if it measures consistently, that is, if it is reliable. Lack of reliability is particularly frequent with tests for young learners, who tend to be influenced by test method features in unpredictable ways (Rea-Dickins 1995). Consequently, tests may be reliable and valid for one group of candidates but not for another. A test is valid if it measures what its developers claim it measures and if it measures what test users think it measures. It follows that validity is not only a quality of the test per se, but also depends on what kind of candidate and what purpose the test is used for. If test develop-

ers' and test users' ideas of what the test measures differ, the test runs the risk of being misused and its results are bound to be misinterpreted. A test of organisational competence which is mistakenly used as one of pragmatic competence will give rise to false claims about the candidates' pragmatic competence. A test, then, may be invalid because it lacks reliability, or because nobody, not even the test developers, know what it measures, or because test users think it measures something different from what test developers know it does. Consequently, any test must be accompanied by adequate information for test users.

Lack of validity leads to unfairness. In order to ensure fairness in testing, principles of test construction, test use and test evaluation should be followed. Such principles are explained in Alderson, Clapham & Wall (1995), where the process of test construction is described and threats to validity which exist at individual stages of test construction are identified. Lack of validity may result from failure to base the test on test specifications, which, in turn, derive from a model of competence. Validity is under threat at the item writing stage, where features may creep into the test which tap abilities that are not intended to be measured by the particular test. Consequently, items should be moderated and pretested before they are used in an actual test. Particularly in the case of rating scales, examiners may misinterpret the scales being used or use them inconsistently. To avoid this, raters must be trained before their ratings are taken as a basis for important decisions. Once tests have been administered and scored or ratings have been carried out, the numerical information is often aggregated into a single score. The way in which scores are aggregated into a single score and how individual subtests or components are weighted must be made explicit and known to test users and test takers in order to avoid misinterpretation of scores. Finally, every test must be checked for validity and the results of these analyses must be made available to test users and test takers in a comprehensible way. Testing institutions adhere to these principles to different degrees, as a recent report of the ILTA Task Force on Testing Standards shows (Alderson, Davidson, Douglas, Huhta, Turner & Wylie 1995). The Association of Language Testers in Europe, for example, have published *The ALTE Code of Practice* (ALTE 1995b), in which the eleven members of ALTE guarantee that the above-mentioned principles will be observed in their tests and examinations and that adequate information for test users and test takers will be provided.

It should become common practice for any test to be analysed for reliability and validity at the piloting stage as well as after it has been administered and used as a basis for decision-making. Clearly, this involves an understanding of the basic concepts of psychometric theory not just by test developers, who carry out the analyses, but also by test users, who have to be able to understand the information provided by test developers (for a survey of approaches to language test validation see, e.g., Alderson, Clapham & Wall 1995:170-188; Sigott 1994).

3 Policy implications

During the last decade language testing has seen considerable theoretical advances. One of the most significant is the emergence of the Bachman model of communicative language ability and test method facets, which is beginning to serve as a theoretical basis for understanding test scores. Test scores are now considered as representing communicative language ability, test method effects, personal attributes and measurement error.

The Bachman model provides a theoretical basis for language testing research. In particular, it enables us to ask relevant questions and formulate hypotheses concerning language test validity. It suggests questions about the extent to which communicative language ability, method effects and test takers' personal attributes are represented in the scores generated by our tests. Tests which are currently in use or being developed must be analysed for their validity now that a more sophisticated theoretical basis is available.

The Bachman model also represents an important step towards a theoretical basis for communicative language testing by providing the concepts necessary for describing communicative tests. The field is now getting ready to incorporate the communicative movement also from a theoretical point of view. We have long given up the narrow view of language as just phonology, lexicon and syntax in language teaching. We should also do so in testing and match up our testing with our teaching.

Although the Bachman model is formulated in English it is abstract enough to serve as the basis for test development in any language. In fact, one of the most urgent needs in Europe is the development of equivalent test batteries in the languages of the European Union. Such a project will necessitate a substantial programme of research and development, in which research in the reduced redundancy paradigm (cloze and C-Testing) will also find its place and have an important part to play. The funding of such a programme would be money well spent.

The structure of my paper, starting with models of language, that is, with the What, and then moving to tests, that is, to the operationalisation of the What, suggests a logical continuation, which is to move from the test to the teaching and, indeed, learning, that is, to the washback of the test. In fact, the question of washback is so important that it would have deserved a section on its own in this paper, if only more research were available. Although everybody seems to be agreed on the existence of washback, there are few empirical studies which examine the effect of language tests on teaching and learning. Notable exceptions are Alderson and Wall (1993), who stake out the ground for research into washback, Wall and Alderson (1993), who report on empirical research into the washback of a new O-level examination in Sri Lanka, as well as Alderson and Hamp-Lyons (ms.), who report research into washback generated by TOEFL in the U.S. All three studies show that teachers', applied linguists' and, indeed, testers' concept of washback has been far too simplistic. At any rate, it seems clear that changing the test does not automatically mean changing the teaching. A test may or may not influence teaching, and, presumably, learning in a variety of ways depending on a variety of factors in the individual teaching situation. The concept of washback has not been properly defined as yet, nor have concrete, testable hypotheses concerning it been formulated until recently. Alderson and Wall (1993) and Alderson and Hamp-Lyons (ms.) therefore propose a set of 15 washback hypotheses. These hypotheses now await empirical study. The issue is complex and urgently requires a substantial programme of research.

References

Alderson, J.C. 1979. The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading, 2,* 108-118.

Alderson, J.C. 1980. Native and nonnative speaker performance on cloze tests. *Language Learning, 1,* 59-76.

Alderson, J.C. 1983. The cloze procedure and proficiency in English as a foreign language. In J.W. Oller (ed.), *Issues in language testing research* (pp. 205-217). Rowley, Mass.: Newbury House.

Alderson, J.C. 1991a. Language testing in the 1990s: How far have we come? How much further have we to go? In S. Anivan (ed.), *Current developments in language testing* (pp. 1-26). Singapore: SEAMEO Regional Language Centre.

Alderson, J.C. 1991b. Bands and scores. In J.C. Alderson & B. North (eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71-86). London: Modern English Publications and The British Council.

Alderson, J.C. & Wall, D. 1993. Does washback exist? *Applied Linguistics, 14*(2), 15-28.

Alderson, J.C., Clapham, C. & Wall, D. 1995. *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, J.C., Davidson, F., Douglas, D., Huhta, A., Turner, C., & Wylie, E. 1995. Report of the Task Force on Testing Standards (TFTS) to the International Language Testing Association (ILTA). July 1995.

Alderson, J.C. & Hamp-Lyons, L. Unpublished Manuscript. TOEFL preparation courses: A study of washback.

Association of Language Testers in Europe. 1995a. *The ALTE Framework.* Cambridge: ALTE.

Association of Language Testers in Europe. 1995b. *The ALTE Code of Practice.* Cambridge: ALTE.

Bachman, L. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. 1991. What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704.

Bachman, L. & Palmer, A. In press. *Language testing in practice: Designing and Developing useful language tests.* Oxford: Oxford University Press.

Beretta, A. 1992. Evaluation of language education: an overview. In J.C. Alderson & A. Beretta (eds.), *Evaluating Second Language Education* (pp. 5-24). Cambridge: Cambridge University Press.

Bialystok, E. 1990. *Communication Strategies.* Oxford: Blackwell.

Bialystok, E. 1991. Achieving proficiency in a second language: a processing description. In R. Phillipson, E. Kellerman, L. Selinker, M. Sharwood Smith & M. Swain (eds.), *Foreign/second language pedagogy research* (pp. 63-78). Clevedon: Multilingual Matters.

Brindley, G. In press. Describing language development? Rating scales and second language acquisition. In L. Bachman & A. Cohen (eds.), *Interfaces between SLA and language testing research.* Cambridge: Cambridge University Press.

Canale, M. 1983. On some dimensions of language proficiency. In J.W. Oller (ed.), *Issues in language testing research* (pp. 333-342). Rowley, Mass.: Newbury House.

Canale, M. & Swain, M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Carroll, J.B. 1961. Fundamental considerations in testing for English proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp.31-40). Washington, D.C.: Center for Applied Linguistics.

Carroll, J.B. 1983. Psychometric theory and language testing. In J.W. Oller (ed.), *Issues in language testing research* (pp. 80-107). Rowley, Mass.: Newbury House.

Cook, G. 1989. *Discourse.* Oxford: Oxford University Press.

Cummins, J. 1979. Cognitive/academic language proficiency, linguistic interdependence, the optimal age question and some other matters. *Working Papers in Bilingualism, 19,* 197-205.

Cummins, J. 1983. Language proficiency and academic achievement. In J.W. Oller (ed.), *Issues in language testing research* (pp. 108-129). Rowley, Mass.: Newbury House.

Cziko, G. 1984. Some problems with empirically-based models of communicative competence. *Applied Linguistics, 5*(1), 23-38.

Davies, A. 1995. Testing communicative language or testing language communicatively: what? how? Paper presented at the British Council Seminar "Communicative language testing revisited", Lancaster, Sept. 11, 1995.

Fulcher, G. 1993. *The construction and validation of rating scales for oral tests in English as a foreign language.* PhD. Thesis. University of Lancaster.

Fulcher, G. Unpublished Manuscript. Does thick description lead to smart tests? A data based approach to rating scale construction. University of Lancaster.

Grotjahn, R. (ed.) 1992. *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Vol. 1. Bochum: Brockmeyer.

Grotjahn, R. (ed.) 1994. *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Vol. 2. Bochum: Brockmeyer.

Grotjahn, R. (ed.) In press. *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Vol. 3. Bochum: Brockmeyer.

Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English.* London: Longman.

Harris, D.P. 1969. *Testing English as a second language.* New York: McGraw Hill.

Heaton, J.B. 1975. *Writing English Language Tests.* London: Longman.

Hymes, D. 1967. Models of the interaction of language and social setting. *Journal of Social Issues, 23*(2), 8-38.

Klein-Braley, Ch. 1981. *Empirical Investigations of cloze tests.* PhD. Thesis. University of Duisburg.

Klein-Braley, Ch. 1994. *Language testing with the C-Test. A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty.* Habilitationsschrift. University of Duisburg.

Köberl, J. & Sigott, G. 1994. Adjusting C-Test difficulty in German. In Grotjahn (1994), pp. 179-192.

Lado, R. 1961. *Language testing.* London: Longman.

Languages Lead Body. 1993. *National Language Standards.* London: Languages Lead Body.

Little, D. 1994. Strategic competence considered in relation to strategic control of the language learning process. Strasbourg: Council of Europe.

McNamara, T.F. 1995. Modelling performance: Opening Pandora's box. *Applied Linguistics, 16*(2), 159-179.

McNamara, T.F. In press. *Measuring second language performance.* London: Longman.

North, B. 1993. Scales of language proficiency. A survey of some existing systems. Strasbourg: Council of Europe.

North, B. 1994. Perspectives on language proficiency and aspects of competence. Strasbourg: Council of Europe.

Oller, J.W. 1979. *Language tests at school.* London: Longman.

Pienemann, M., Johnston, M. & Brindley, G. 1988. Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition, 10,* 217-243.

Pollitt, A. 1991. Giving students a sporting chance: Assessment by counting and by judging. In J.C. Alderson & B. North (eds.), *Language testing in the 1990s: The communicative legacy* (pp. 46-59). London: Modern English Publications and The British Council.

Raatz, U. & Klein-Braley, Ch. 1985. How to develop a C-Test. *Fremdsprachen und Hochschule, 13/14,* 20-22.

Rea-Dickins, P. 1991. What makes a grammar test communicative? In J.C. Alderson & B. North (eds.), *Language testing in the 1990s: The communicative legacy* (pp. 112-131). London: Modern English Publications and The British Council.

Rea-Dickins, P. 1995. To test or not to test: Issues in the assessment of young children learning EFL. Paper presented at the British Council Seminar "Communicative language testing revisited", Lancaster, Sept. 7, 1995.

Sang, F., Schmitz, B., Vollmer, H.J., Baumert, J. & Roeder, P.M. (1986). Models of second language competence: a structural equation approach. *Language Testing, 3*(1), 54-79.

Sigott, G. 1992. Problems in the development of a proficiency measure with crosslinguistic applicability. *Arbeiten aus Anglistik und Amerikanistik, 17*(2), 151-170.

Sigott, G. 1993. *Zur Lernbarkeit von Englisch und Französisch für deutsche Muttersprachler. Eine explorative Pilotstudie.* Tübingen: Narr.

Sigott, G. 1994. Language test validity: An overview and appraisal. *Arbeiten aus Anglistik und Amerikanistik, 19*(2), 287-294.

Sigott, G. In preparation. *Reduced redundancy language testing. A construct identification study of cloze and C-Tests.*

Sigott, G. & Köberl, J. In press. Deletion patterns and C-Test difficulty across languages. In Grotjahn (In press).

Skehan, P. 1988. State of the art article: Language Testing I. *Language Teaching, 21*(4), 211-221.

Skehan, P. 1989. State of the art article: Language Testing II. *Language Teaching, 22*(1), 1-13.

Skehan, P. 1991. Progress in language testing. The 1990s. In J.C. Alderson & B. North (eds.), *Language testing in the 1990s: The communicative legacy* (pp. 3-21). London: Modern English Publications and The British Council.

Spolsky, B. 1995. *Measured Words.* Oxford: Oxford University Press.

van Ek, J.A. 1975. *Threshold Level.* Strasbourg: Council of Europe.

Wall, D. & Alderson, J.C. 1993. Examining washback: the Sri Lankan Impact Study. *Language Testing, 10*(1), 41-69.

Weir, C. 1995. The communicative testing of reading. Paper presented at the British Council Seminar "Communicative language testing revisited", Lancaster, Sept. 8, 1995.

Widdowson, H.G. 1979. *Explorations in applied linguistics.* Oxford: Oxford University Press.

Topics for work groups

1 Communicative testing: Practice and problems

The group will be shown three videorecordings of authentic oral exams. Participants will be asked to decide on a rating procedure and apply it. This should lead to a discussion of theoretical and practical issues.

2 Test equivalency across languages

The group could be given part of an English proficiency test and be asked to translate it into another language. This should lead to an understanding of some of the problems involved.

3 Washback: A closer look

The group will be presented with Alderson et al.'s washback hypotheses. If necessary, as a preliminary step the hypotheses will be clarified. Participants will then be asked to formulate concrete hypotheses for their individual situation in their home country. This should generate an awareness of the need for research into washback.

4 Reduced redundancy testing: an introduction

Participants can be asked to take a C-Test. Then they would be asked what they think the abilities were they have been tested on. This should lead to a discussion of basic assumptions of C-Testing and strengths and weaknesses of the approach.